Vertrauenswürdige Systeme mit künstlicher Intelligenz

Nur wenn wir Systemen vertrauen, die auf künstlicher Intelligenz (KI) basieren, können wir das ganzheitliche Potenzial der KI für uns als Individuen, für Organisationen und Unternehmen bzw. für unsere Gesellschaft nachhaltig nutzen. Zentrale Aspekte der Vertrauenswürdigkeit von KI-Systemen können durch die Anwendung der Distributed Ledger Technology hergestellt werden.

Künstliche Intelligenz hat in vielen Bereichen das Potenzial, Prozesse und Entscheidungsfindungen im wahrsten Sinne zu automatisieren. Anwendungen finden sich etwa in der Industrie 4.0, dem autonomen Fahren oder im Gesundheitswesen. In Letzterem können KI-basierte Systeme heutzutage beispielsweise Echokardiogramme schneller und genauer analysieren als medizinische Experten. Trotz aller Erfolge, die das zukünftige Potenzial von KI-Anwendungen demonstrieren, müssen stets auch das Vertrauen in KI-basierte Systeme und die Implikationen ihrer Anwendung berücksichtigt werden. Ein Aspekt der Vertrauenswürdigkeit ist beispielsweise die Tatsache, dass das Training von KI-Modellen viele hochqualitative Daten, welche durchaus sehr sensibel sein können, benötigt. Ein weiterer Aspekt der Vertrauenswürdigkeit ist die Erklärbarkeit von KI-Modellen selbst. Häufig ist die Funktionsweise von KI-Modellen intransparent und schwer nachzuvollziehen. Diese Aspekte spielen insbesondere bei der Anwendung von KI in kritischen Informationsinfrastrukturen, wie z. B. der Anwendung von KI im Gesundheitswesen, eine zentrale Rolle: Eine erste Herausforderung ist es, hochqualitative Daten in ausreichender Form zu erlangen und sicherzustellen, dass diese nach höchsten Datenschutzstandards verarbeitet werden. Selbst wenn ein KI-Modell erfolgreich trainiert wurde, stellt die Erklärbarkeit dieses Modells eine weitere Herausforderung dar: Wenn ein KI-Modell zum Beispiel eine bestimmte Therapie empfiehlt, aber weder der Patient noch der Arzt oder ein Informatikexperte nachvollziehen können, wie diese Empfehlung zustande gekommen ist, fehlen Arzt und Patient gleichermaßen das Vertrauen in das KI-System

Prof. Dr. Ali Sunyaev *(✉)*
ist Professor für Informatik und Direktor am Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB) des Karlsruher Instituts für Technologie (KIT). Seine Forschungsinteressen sind zuverlässige Internettechnologien und komplexe Anwendungen im Gesundheitswesen. In seiner wissenschaftlichen Arbeit untersucht Ali Sunyaev die vielschichtigen Nutzungskontexte von internetbasierten Systemen, deren Entwicklung und Pilotierung in realen Anwendungsszenarien und die wechselseitigen Wirkungszusammenhänge zwischen menschlichem Verhalten und Informationstechnologie. Ali Sunyaev wurde für seine Forschung mehrfach ausgezeichnet und fungiert als Mentor für zahlreiche Start-ups.
sunyaev@kit.edu

Institut für Angewandte Informatik und Formale Beschreibungsverfahren, Karlsruher Instituts für Technologie, Karlsruhe, Deutschland

Wirtschaftsinformatik & Management 2020 • 12 (2): 111–113
https://doi.org/10.1365/s35764-020-00250-6
Online publiziert: 5. März 2020

und sie könnten daher von der Anwendung der KI-basierten Empfehlung absehen.

Distributed Ledger Technology (DLT) hat das Potenzial, die Vertrauenswürdigkeit von KI-basierten Systemen zu erhöhen und als Schlüsseltechnologie deren Bereitstellung in kritischen Informationsinfrastrukturen, wie der digitalisierten Medizin, zu ermöglichen: DLT erzeugt in erster Linie einen Konsensus zwischen den Knoten eines Peer-to-Peer-basierten Ledgers, obwohl manche Knoten möglicherweise vorübergehend nicht verfügbar sind oder einige Knoten versuchen, betrügerische oder falsche Daten aufzunehmen. Also ermöglicht DLT, die sogenannten „byzantinischen Fehler" zu tolerieren und wendet spieltheoretische Konzepte an, um die beschriebenen Problematiken tolerieren zu können (z. B. das Problem der byzantinischen Generäle). Des Weiteren verhindert eine Dezentralisierung des Datenmanagements weitere Unsicherheiten, die in zentralisierten Infrastrukturen auftreten, zum Beispiel das, was technisch als „Single Point of Failure" bezeichnet wird. Fällt ein zentrales System aus (der sogenannte „Point of Failure"), auf das viele andere Dienste zugreifen und von dem sie abhängen, ist nicht nur das eigentliche System betroffen, sondern auch alle anderen Dienste. Die DLT ermöglicht es, nicht von einer zentralen Instanz abhängig zu sein. Das geläufigste DLT-Konzept ist die Blockchain, welche zum Beispiel durch die darauf aufbauende Kryptowährung Bitcoin bekannt wurde. Diese Technologie ermöglicht aber weitere Konzepte und Implementierungen, die jeweils inhärente Trade-offs in Bezug auf verschiedene Charakteristika aufweisen (zum Beispiel Privatsphäre der Daten, Transaktionsfinalität, Sicherheit gegen Attacken) [1].

Ein DLT-basiertes System im Gesundheitswesen kann beispielsweise dergestalt aussehen, dass Patienten über eine Anwendung ihre Daten mittels eines im Hintergrund laufenden Distributed Ledgers verwalten. Insbesondere können die Patienten dabei weiteren Beteiligten Zugriff auf die Daten geben, beispielsweise einem weiteren Arzt zur Einholung einer ärztlichen Zweitmeinung oder auch einem Forscherteam. Die Patienten können steuern und jederzeit einsehen, wer wann auf welche Daten zugreifen darf bzw. zugegriffen hat, und womöglich auch, wozu diese Daten verwendet werden oder wurden. Technisch können die Daten entweder in verschlüsselter Form direkt auf dem Distributed Ledger gespeichert werden, oder – da praktikabler – außerhalb des Distributed Ledgers gespeichert werden mittels Verwaltung der Zugriffsrechte durch DLT.

Insbesondere kann ein solches System auch als Grundlage für ein DLT-basiertes föderiertes Lernen verwendet werden. Dabei trainieren verschiedene Einheiten KI-Modelle mit Daten lokal. Die trainierten KI-Modelle werden über den Distributed Ledger untereinander ausgetauscht und zu einem neuen Modell zusammengeführt. Das neue KI-Modell profitiert von den Daten aller Einheiten, wohingegen die einzelnen Einheiten lediglich KI-Modelle nach außen senden und nicht die privaten Trainingsdaten an sich. Über DLT-basierte Kryptowährungen können weiterhin Anreize z. B. dergestalt geschaffen werden, dass Beteiligte finanziell belohnt werden, sofern sie neue KI-Modelle teilen. Dies erhöht für Nutzer die Nachvollziehbarkeit, welche Daten in die Erstellung des KI-basierten Systems eingeflossen sind, und wie das KI-System darauf basierend Entscheidungen trifft. Durch ein solches System kann gleichermaßen die Verfügbarkeit von Daten für KI-Systeme wie auch die Vertrauenswürdigkeit von Nutzern in diese KI-Systeme gestärkt werden. Hinzu können durch einen transparenten Distributed Ledger, der Daten speichert, KI-Experten, Zertifizierungsexperten, sowie versierte Nutzer die Funktionsweise des Systems nachvollziehen, welches das Vertrauen aller Nutzer in das System stärken kann.

Die Anwendungsfälle für vertrauenswürdige KI sind vielfältig: Neben dem hier aufgeführten Beispiel im Gesundheitswesen können zum Beispiel in der Industrie 4.0 verschiedene Fabriken Daten sicher miteinander teilen oder autonome Fahrzeuge voneinander lernen, ohne dass die Hoheit über die Daten in die Hände einiger weniger fällt. Zusammenfassend lässt sich zum einen sagen, dass erst vertrauenswürdige KI die vollen Potenziale der KI entfalten kann. Zum anderen, dass DLT eine vielversprechende Technologie ist, diese notwendige Vertrauenswürdigkeit in KI-Systeme herzustellen. Dies aber erfordert noch weitere, interdisziplinäre Forschungsarbeit an der Schnittmenge von KI und DLT, welche vielversprechende Integrationsmöglichkeiten bietet [2].

Funding. Open Access funding provided by Projekt DEAL.

Literatur

[1] Kannengießer, N., Lins, S., Dehling, T., & Sunyaev, A. (2020). Trade-offs between distributed ledger technology characteristics. *ACM Computing Surveys*, 1–35. arXiv preprint arXiv:1906.00861.

[2] Pandl, K. D., Thiebes, S., Schmidt-Kraepelin, M., & Sunyaev, A. (2020). On the convergence of artificial intelligence and distributed ledger technology: a scoping review and future research agenda. 1–24. arXiv preprint arXiv:2001.11017.